Human Post-editing in Hybrid Machine Translation Systems: Automatic and Manual Analysis and Evaluation

Chapter · March 2018

DOI: 10.1007/978-3-319-77703-0_26

CITATIONS
4

READS
543

3 authors:

Juncal Gutiérrez-Artacho
University of Granada
68 PUBLICATIONS **246** CITATIONS
SEE PROFILE

María Dolores Olvera-Lobo
University of Granada
174 PUBLICATIONS **848** CITATIONS
SEE PROFILE

Irene Rivera Trigueros
University of Granada
24 PUBLICATIONS **22** CITATIONS
SEE PROFILE

This is a pre-copyedited version of a contribution published in Rocha Á., Adeli H., Reis L.P., Costanzo S. (eds) Trends and Advances in Information Systems and Technologies. WorldCIST'18 2018. Advances in Intelligent Systems and Computing, vol 745 published by Springer. The definitive authenticated version is available online via https://doi.org/10.1007/978-3-319-77703-0_26

Human Post-editing in Hybrid Machine Translation Systems: Automatic and Manual Analysis and Evaluation

Juncal Gutiérrez-Artacho[1][0000-0002-0275-600X], María-Dolores Olvera-Lobo[2,3][0000-0002-0489-7674] and Irene Rivera-Trigueros[4] (✉)[0000-0003-4877-4083]

[1] University of Granada, Department of Translation and Interpreting, Faculty of Translation and Interpreting, C/ Buensuceso, 11, 18003, Granada, Spain
juncalgutierrez@ugr.es
[2] University of Granada, Department of Information and Communication, Colegio Máximo de Cartuja, Campus Cartuja s/n, 18071, Granada, Spain
[3] CSIC, Unidad Asociada Grupo SCImago, Madrid, Spain
molvera@ugr.es
[4] University of Granada
irerivera@correo.ugr.es

Abstract. This study assesses, automatically and manually, the performance of two hybrid machine translation (HMT) systems, via a text corpus of questions in the Spanish and English languages. The results show that human evaluation metrics are more reliable when evaluating HMT performance. Further, there is evidence that MT can streamline the translation process for specific types of texts, such as questions; however, it does not yet rival the quality of human translations, to which post-editing is key in this process.

Keywords: Hybrid Machine Translation, automatic evaluation, human evaluation, post-editing,

1 Introduction

Access to information is increasingly global, which brings with it the growth in a non-English speaking public and, as such, a demand for tools that allow users to access this information. Faced with this scenario, when assessing an IR (Information Retrieval) system we find, among a number of key aspects, its capacity for enabling users to find a corpus of documents in different languages, and provide the relevant information despite limitations of linguistic competence [1].

So-called CLIR (Cross Language Information Retrieval) systems retrieve relevant documents without affording importance to the language of the query [2]. The fact that these systems involve the participation of at least two languages makes it necessary to apply a translation tool. Machine Translation (MT) is one of the most utilised tools by

these systems to carry out translative processes [3–6]. Nevertheless, the majority of researchers agree that MT systems are not developed enough to overcome the barriers that language poses in CLIR systems [5].

At present, ever more search applications have been focusing on queries formulated in natural language [7]. It is therefore interesting to analyse MT performance regarding the translation of questions made in natural language, from the perspective of its potential as a CLIR tool.

The translation market can be classified as global, decentralised, specialised, dynamic, virtual and demanding [8]. The incorporation of information and communication technologies has been pivotal for the development of new tools to help professionals in this field. This is the case for MT systems, which can be integrated together with other resources to carry out the translation process more quickly and efficiently [9, 10]. Although MT does not boast the level of excellence of human translations [9, 11], it is useful in the development of this process.

Globalisation and companies' intentions to expand towards international markets has meant an increase in MT use, as in many cases it is impossible to satisfy translation demand while also seeking to reduce costs as far as possible [12]. Another factor having an impact on this field are expectations regarding demands for quality; on occasions it is enough for a general idea about the content to satisfy the needs of the client [11]. All of these changes have made post-editing—the revision process for a text that has been previously translated by an MT system—increasingly important. This process is carried out to correct possible errors, adjusting itself to quality criteria and making the least edits possible [10].

Furthermore, measuring the performance of an MT system is essential to be able to progress in its research, development and improvement. However, assessing the MT can present difficulties as in the majority of cases there is not just one correct translation [13, 14]. This is why there are a number of metrics and criteria for undertaking the evaluation of MT systems.

The main objective of this study is the evaluation, both automatic and manual, of the performance of two hybrid machine translation systems (HMT) via a corpus of questions used for IR system evaluation. This process will allow us, as well as taking a closer look at new trends in MT and IR, to assess whether the automatic evaluation metrics are sufficient to determine the quality of an MT system or whether, on the contrary, it is necessary to combine these metrics with human evaluation to obtain more reliable results.

1.1 New trends in Machine Translation

Given the continuous evolution of technology, MT can be understood from different perspectives. One of these perspectives emphasises the complete or partial automatization of the translation of one natural language to another [15, 16]. One of the current trends in MT is the combination of different types of methods, giving rise to hybrid technologies [17, 18]. These new systems combine the advantages of two different approaches: rule-based MT and statistical or analogy-based MT. Thus, there is an attempt at solving problems detected in these types of technologies to produce translations with

greater accuracy and quality [19, 20]. Hybridization can be carried out via a single engine—*Single-Engine Hybridization* (SEH)—or various—*Multi-Engine Hybridization* (MEH)[21].

1.2 Evaluation of machine translation

The combination of various metrics that evaluate different aspects can lead us to more reliable results. There are two main types of MT evaluation: human and automatic. Human evaluation, undertaken by experts, is more reliable, but more expensive, requires more time and is more subjective. Human evaluation revolves around the categories of fluency and adequacy [22].

The metrics requiring human intervention include SSER (Subjective Sentence Error Rate), where each segment is evaluated according to an error scale between 0 and 10, taking into account both adequacy and fluency [23]. In this case, being translations of short questions, a scale between 0 and 3 was employed to simplify the manual evaluation process.

0: unintelligible
1: comprehension difficult (serious syntax and/or content errors)
2: comprehension acceptable (minor syntax and/or content errors)
3: correct both in terms of syntax and semantics

On the other hand, machine evaluation reduces both costs and time necessary to carry out the evaluation, with just an algorithm being necessary for it to work, which guarantees objectivity [13]. However, the values obtained with this type of evaluation are artificial and a high value does not necessarily imply the quality of a translation [22].

In general, the metric most used for evaluating MT quality is BLEU [24]. Notwithstanding, it is criticised due to the difficulty in interpreting its results, and for not measuring the quality of translations, rather their similarity with reference translations [25–27]. To attempt to obtain the most reliable results possible, seven more metrics are applied, in addition to BLEU

BLEU (Bilingual Evaluation Understudy). BLEU is a precision metric carried out at the level of n-grams, indivisible linguistic units. A modified precision is used that takes into account the maximum number of appearances of each n-gram in the reference translation, and a brevity penalty is applied, which is added to the calculation of the metric [28].

$$BLEU = BP \cdot \exp(\sum_{n=1}^{N} W_n \log p_n) \quad (1)$$

GTM (General Text Matcher). GTM calculates the precision, exhaustiveness and f-measure measure, based on the maximum number of unigrams that coincide. This metric favours coincidences that are longer and in the correct order, as it assigns them a greater weighting in the calculation of the metric parameters [29]. GTM has variants

that depend on the weighting it assigns to the longest coincidences. The GTM-3 variant has been selected for this study.

METEOR (Metric for Evaluation of Translation with Explicit ORdering). METEOR is based on the word-for-word correspondence between the MT generated translation and one or more reference translations. Correspondence is not just made between identical words, but also words with the same root and synonyms, for which it employs different modules. The METEOR-ex variant has been used for this study, with machine assessment carried out, initially employing the *exact* module, which associates two unigrams if they are exactly the same[30].

$$METEOR = Fmean * (1 - Penalty) \quad (2)$$

ROUGE (Recall Oriented Understudy for Gisting Evaluation). A metric very similar to GTM and METEOR, as it is also based on precision, exhaustiveness and, to an extent, symmetry, for MT evaluation [31]. ROUGE, however, does not apply the brevity penalty. In this case, the variant ROUGE-L has been selected to carry out this investigation, which takes into account the length of the longest sequences which coincide between the candidate translation and the reference translation, to undertake the evaluation [13, 32].

WER (Word Error Rate). WER is based on the Levenhstein distance, or editing distance. This metric does not admit the reordering of word and so substitutions, eliminations and insertions incur the same penalty. The number of editing operations are divided between the number of words from the reference translation [14, 33].

$$WER\ (p) = \frac{1}{N^*_{ref}} \sum_{k=1}^{K} \min_r d_L\ (ref_{k,r}, hyp_k) \quad (3)$$

PER (Position-independent word error rate). PER attempts to solve the problem created by WER, by not allowing word reordering. This metric compares the words in the two sentences without taking order into consideration [14, 33].

$$PER\ (p) = \frac{1}{N^*_{ref}} \sum_{k=1}^{K} \min_r d_{PER}\ (ref_{k,r}, hyp_k) \quad (4)$$

TER (Translation Error Rate). TER counts the number of edits required for an MT generated translation to coincide with a reference translation. This metric allows the reordering of words and furthermore considers it as one more edit together with insertions, eliminations and replacements [14].

$$TER = \frac{\#\ of\ edits}{average\ \#\ of\ reference\ words} \quad (5)$$

2 Methodology

To respond to the objectives of the investigation, the machine translators used in the study had to be free of charge, have Spanish and English amongst the available languages, and apply hybrid technology. Systran and ProMT were the only HMTs to fulfil all of the established requirements. In 2009, Systran launched the first hybrid MT motor onto the market. For its part, ProMT, like Systran, began as a rule-based system, but in recent years the company has created an HMT system through the incorporation of statistical techniques [34].

The text corpus used is comprised of a collection of evaluation questions proposed by the CLEF (Cross Language Evaluation Forum). These collections are used on this type of forum to carry out the evaluation of techniques and IR systems, allowing comparative studies to be performed [3, 35–44]. In order to carry out the study presented here we used two collections of questions on European legislation from the ResPubliQA track (2009 and 2010), related to the Europarl corpus, which includes the minutes of the European Parliament in a number of languages [45]. The collection of questions is available in various languages and they have been translated by human translators. The corpus, comprised of a sample of 100 questions, was translated from English (EN) into Spanish (ES), and from Spanish into English, both with ProMT and Systran. This gave the result of a total of 400 translations—200 EN-ES and 200 ES-EN. Later, the translations were evaluated both automatically and manually. The automatic evaluation was carried out with ASIYA[1], a tool developed by the Polytechnic University of Catalonia, which allows machine generated translations to be assessed.

Firstly, an evaluation was made of each question translated by Systran and ProMT individually, both EN-ES and ES-EN. Given that the corpus questions are translated into various languages by human translators, it was possible to employ them as reference translations when comparing translations generated by machine translators. The metrics employed for automatic evaluation were the aforementioned—BLEU, METEOR-ex, ROUGE-L, TER, WER, PER and GTM-3. The manual evaluation was undertaken based on the criteria determined by the SSER metric.

3 Results

3.1 Automatic evaluation

The results obtained by both MT systems following evaluation with ASIYA are quite similar (Fig.1 and Fig. 2). ProMT is the machine translator that gives the best results both when translating from English to Spanish and vice-versa.

[1] Available at http://asiya.cs.upc.edu/ (Last visit 05/01/2018)

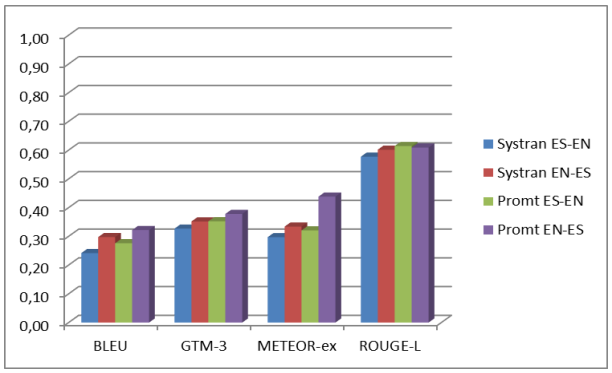

Fig. 1. BLEU, GTM-3, METEOR-ex and ROUGE-L average results

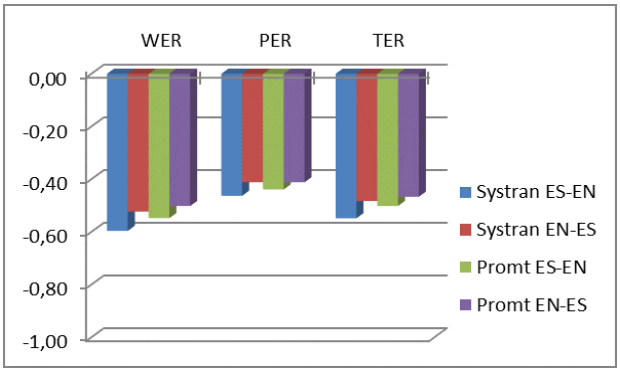

Fig. 2. WER, PER and TER average results.

Basing ourselves on the results obtained following the automatic evaluation, if we analyse the values obtained by both translators as a whole, it could be stated that the general performance of both translators is not adequate for producing quality translations (Table 1). Save for the values of ROUGE-L and PER, none of the values exceeds half of the maximum value of each metric.

Table 1. General performance of both MT systems

Metrics	Total average	Maximum value
BLEU	0.28	1
GTM-3	0.35	1
WER	-0.54	0
PER	-0.43	0
METEOR-ex	0.35	1
ROUGE-L	0.60	1
TER	-0.50	0

3.2 Manual evaluation

ProMT is a machine translator that obtains better results with manual evaluation (Fig. 3), especially when translating ES-EN: nearly half of the translations generated (49%) only contain minor errors relating to content or syntax.

On the other hand, Systran performs the poorest with ES-EN translation: over a third of the translations generated are unintelligible (38%).

For the general performance of both machine translators, translations with an acceptable comprehension stand out (39%), which show minor syntax or content errors; therefore, these phrases would be quick and simply to post-edit. In second place (23.5%) are those translations that can be understood, but with great difficulty, due to presenting serious grammar or content errors. These translations could be post-edited but would require more time and effort. The next group is that of unintelligible translations (22.25%); it is not worth post-editing these translations, as this would take longer than human translation. Correct translations in terms of syntax and semantics, that is, those which would not need post-editing, is the least numerous group; despite this, it can be considered as an acceptable result, given that 15% of translations would not require human intervention.

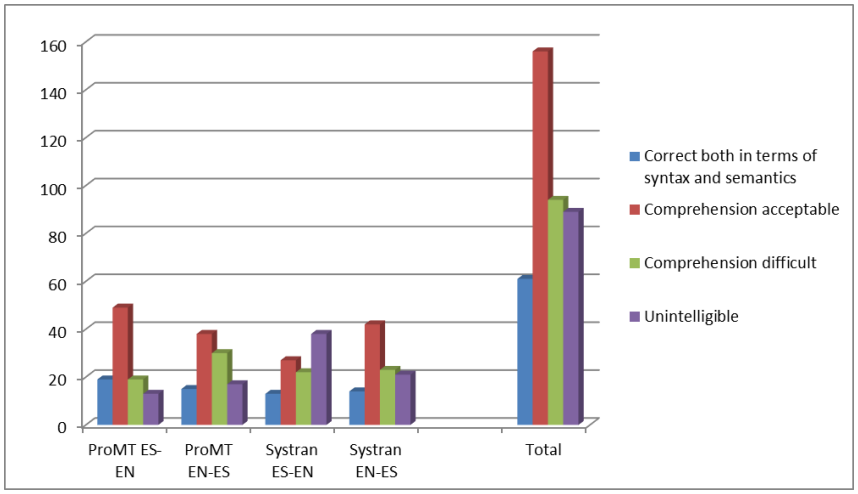

Fig. 3. Results of the manual evaluation.

4 Conclusions

The analysis carried out allows us to verify that automatic evaluation in the majority of cases is insufficient for assessing MT performance. It is always recommendable to also carry out human evaluation as it may better determine translation quality. Notwithstanding, it is important to define the criteria with which human evaluation must be implemented in order for it to be as accurate as possible, always taking into account that it will have a subjective component.

The carrying out of both human and automatic evaluation of the HMTs Systran and ProMt has allowed us to confirm that, although automatic evaluation seemed not to perform well following the analysis of the results, after the human evaluation it has been observed that the performance was better than that indicated by the automatic evaluation metrics, as approximately 15% of the translations were correct in both semantics and syntax, and around 40% of the translations only had minor errors, to which little time needs to be employed on their post-editing.

In this case, MT has accelerated the process for the translation of a corpus of 400 short questions created for their input into IR systems. As we have seen, MT, although not reaching the quality of human translation, can be employed for time-saving purposes, above all when it involves a large volume of short translations that are not too difficult. However, MT cannot be appropriate for other types of longer or more complex texts. Even if MT can aid the translation process, we must not forget that machine translators are not yet capable of matching the quality of human translations, to which post-editing is becoming a new stage in the translation process.

Acknowledgements. This work was supported by the University of Granada Special Research Programme - Starting Research Grants for Master Students.

References

1. Gutiérrez Artacho, J.: Recursos y herramientas lingüísticos para los sistemas de búsqueda de respuestas monolingües y multilingües, (2015).
2. Zhou, D., Truran, M., Brailsford, T., Wade, V., Ashman, H.: Translation Techniques in Cross-language Information Retrieval. ACM Comput. Surv. 45, 1, 1–1:44 (2012).
3. Olvera-Lobo, M., Gutierrez-Artacho, J.: Language resources used in multi-lingual question-answering systems. Online Inf. Rev. 35, 543–557 (2011).
4. Olvera-Lobo, M.D., Garcia-Santiago, L.: Analysis of errors in the automatic translation of questions for translingual QA systems. J. Doc. 66, 434–455 (2010).
5. García-Santiago, L., Olvera-Lobo, M.-D.: Analysis of automatic translation of questions for question-answering systems. Inf. Res. 15, (2010).
6. Madankar, M., Chandak, M.B., Chavhan, N.: Information Retrieval System and Machine Translation: A Review. Procedia Comput. Sci. 78, 845–850 (2016).
7. Gupta, M., Bendersky, M.: Information Retrieval with Verbose Queries. Found. Trends Inf. Retriev 9 (3-4), 200–354 (2015).
8. Olvera-Lobo, M.D., Castro-Prieto, M.R., Quero-Gervilla, E., Munoz-Martin, R., Munoz-Raya, E., Murillo-Melero, M., Robinson, B., Senso-Ruiz, A., Vargas-Quesada, B., Dominguez-Lopez, C., int, A., Granada, U.: Translator training and modern market demands. Perspect. Transl. 13, 132–142 (2005).
9. Koponen, M.: Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. J. Spec. Transl. 131–148 (2016).
10. Mesa-lao, B.: Introduction to post-editing – The CasMaCat GUI 1 (2013).
11. Allen, J.: Post-editing. In: Somers, H.L. (ed.) Computers and Translation: a translators guide, 297–317. John Benjamins, Amsterdam/Philadelphia (2003).

12. Lagarda, A.L., Ortiz-Martinez, D., Alabau, V., Casacuberta, F.: Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Comput. SPEECH Lang. 32, 109–134 (2015).
13. Shaw, F., Gros, X.: Survey of Machine Translation Evaluation. Saarbrücken (2007).
14. Mauser, A., Hasan, S., Ney, H.: Automatic Evaluation Measures for Statistical Machine Translation System Optimization. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). Marraketch, Morocco (2008).
15. Hutchins, W.J., Somers, H.L.: An Introduction to Machine Translation. Academic Press, London. 57, 377 (1992).
16. Arnold, D., Balkan, L., Meijer, S., Humphreys, R.L., Sadler, L.: An Introductory Guide. NCC Blackwell, London (1994).
17. Costa-Jussa, M.R., Fonollosa, J.A.R.: Latest trends in hybrid machine translation and its applications. Comput. SPEECH Lang. 32, 3–10 (2015).
18. Labaka, G., España-Bonet, C., Màrquez, L., Sarasola, K.: A hybrid machine translation architecture guided by syntax. Mach. Transl. 28, 91–125 (2014).
19. Hunsicker, S., Yu, C., Federmann, C.: Machine Learning for Hybrid Machine Translation. In: Proceedings of the Seventh Workshop on Statistical Machine Translation, 312–316. Montreal (2012).
20. Tambouratzis, G.: Conditional random fields versus template-matching in MT phrasing tasks involving sparse training data. PATTERN Recognit. Lett. 53, 44–52 (2015).
21. Kamran, A.: Hybrid Machine Translation Panel (2013).
22. Leusch, G.: Evaluation Measures in Machine Translation (2005).
23. Niessen, S., Och, F.J., Leusch, G., Ney, H.: An evaluation tool for machine translation: Fast evaluation for MT research. ACM Trans. Inf. Syst. 20, 39–45 (2000).
24. Mayor, A., Alegria, I., Díaz Ilarraza, A., Labaka, G., Lersundi, M., Sarasola, K.: Evaluación de un sistema de traducción automática basado en reglas o por qué BLEU sólo sirve para lo que sirve. Proces. leng. nat. 43, 197-205 (2009).
25. Homola, P., Kubon, V., Pecina, P., Pecina, P.: A Simple Automatic MT Evaluation Metric. In: Association for Computational Linguistics (ed.) Proceedings of the 4th EACL Workshop on Statistical Machine Translation, 33–36. Athens (2009).
26. Turian, J.P., Shen, L., Melamed, I.D.: Evaluation of Machine Translation and its Evaluation. In: Proceedings of MT Summit IX. New Orleans (2003).
27. Boitet, C., Bey, Y., Tomokiyo, M., Cao, W., Blanchon, H.: IWSLT-06 : experiments with commercial MT systems and lessons from subjective evaluations (2006).
28. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 311–318. Association for Computational Linguistics, Stroudsburg, PA, USA (2002).
29. Bouillon, P. (Coord.): Analysis of existing metrics and proposal for a task-oriented metric. In: ACCEPT- Automated Community Content Editing PorTal (2012).
30. Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 65–72. Ann Arbor (2005).
31. Goutte, C.: Learning Machine Translation. MIT Press, Cambridge/Mass (2009).
32. Ferrández, O.; Micol, D.; Muñoz, R.; Palomar, M.: Técnicas léxico- sintácticas para el reconocimiento de Implicación Textual. Proces. del Leng. Nat. 38, 53–60 (2007).

33. Popovic, M., Ney, H.: Word Error Rates: Decomposition over POS classes and Applications for Error Analysis. In: Proceedings of the Second Workshop on Statistical Machine Translation, 48–55. Association for Computational Linguistics, Prague (2007).
34. Helle, A.: Hibridación en lenguas distantes, (2013).
35. Olvera-Lobo, M.D.; Quero-Gervilla, E.; Robinson, B.; Senso-Ruiz, J.A.. C., M.R.; Muñoz-Martín, R.; Muñoz-Raya, E. y Murillo-Melero, M: Presentation of a distance training model for introduction into the practice of teaching translation according to the requirements of the Bologna declaration. Vestn. MGU. Ser. nº 26. Pedagog. Sci. 605, 196–208 (2010).
36. Olvera-Lobo, M.-D., Gutiérrez-Artacho, J.: Evaluación de los sistemas QA de dominio abierto frente a los de dominio especializado en el ámbito biomédico. In: I Congreso Español de Recuperación de Información (CERI 2010), 161–169. Madrid (2010).
37. Olvera-Lobo, M.-D., Gutierrez-Artacho, J.: Open- vs. Restricted-Domain QA Systems in the Biomedical Field. J. Inf. Sci. 37, 152–162 (2011).
38. Olvera-Lobo, M.-D., Gutierrez-Artacho, J.: Multilingual Question-Answering System in Biomedical Domain on the Web: An Evaluation. In: Forner, P and Gonzalo, J and Kekalainen, J and Lalmas, M and DeRijke, M (ed.) Multilingual and Multimodal Information Access Evaluation, 83-88 (2011).
39. Olvera-Lobo, M.-D., Gutierrez-Artacho, J.: Performance Analysis in Web-based Question Answering Systems. Rev. Esp. Doc. Cient. 36, (2013).
40. Olvera-Lobo, M.D., Gutiérrez-Artacho, J.: Searching Health Information in Question-Answering Systems. In: Cruz-Cunha, M.M., Miranda, I.M., and Gonçalves, P. (eds.) Handbook of Research on ICTs for Human-Centered Healthcare and Social Care Services., 474–490. IGI Global, Hershey (2013).
41. Olvera-Lobo, M.D., Gutiérrez-Artacho, J.: Nuevas tendencias en recuperación de información: la búsqueda de respuestas desde la perspectiva de la traducción. In: VI Congreso Internacional de la Asociación Ibérica de Estudios de Traducción e Interpretación: 'Traducimos desde el sur'. Asociación Ibérica de Estudios de Traducción e Interpretación (AIETI), Las Palmas de Gran Canaria (2013).
42. Olvera-Lobo, M.D., Gutiérrez-Artacho, J.: Satisfacción de los usuarios en la búsqueda multilingüe de respuestas como recursos de información terminológica. In: Vargas-Sierra, C. (ed.) TIC, trabajo colaborativo e interacción en terminología y traducción, 191–200. Comares, Granada (2014).
43. Olvera-Lobo, M.D., Gutiérrez-Artacho, J.: Overview of Translation Techniques in Cross-Language Question Answering during the Last Decade. In: Khosrow-Pour, M. (ed.) Encyclopedia of Information Science and Technology, 4747–4755. IGI Global, Hershey (2015).
44. Olvera-Lobo, M.-D., Gutierrez-Artacho, J.: Question Answering Track Evaluation in TREC, CLEF and NTCIR. In: Rocha, A and Correia, AM and Costanzo, S and Reis, LP (ed.) New Contributions in Information Systems and Technologies, 1, 13–22 (2015).
45. Koehn, P.: Europarl : A Parallel Corpus for Statistical Machine Translation. MT Summit. 11, 79–86 (2005).